Course Information

Lecturer: Dr. Rashed Iqbal
Office: Bunche 2253
Time and Location: Saturday 9:00AM – 12:00 PM, Dodd Hall #78
Office Hours: Saturday, 12:00 PM to 1:00 PM
Telephone: (805) 341-8450
E-mail: rashed_iqbal@econ.ucla.edu

Course Description

Data Analytics and Big Data have found applications in every domain of knowledge where data exists including finance, business, sciences, and humanities. The domain touches us daily when we get a movie recommendation, when we use a navigation app to find our way, or when we search the web. Employers thus seek professionals called Data Scientists that collect, visualize and analyze data for gaining actionable insights and for forecasting and prediction using sophisticated machine learning and statistical algorithms. The requirements for the many of these skills however are no longer limited to Data Scientists. Today almost all professionals need to work with data in the modern context and economists are no exceptions.

This course focuses to provide you with a solid basis for working with data and for exploring the discipline. You will learn collection, visualization, analysis, and processing of Big Data through lectures, case studies, and via an intensive class project. Tableau and Python will be used in the class. The course will address both the theoretical underpinning of the domain as well as intensive applied computing component. This is an intense program and you will be required to work additional hours to cover any gaps in computing, mathematics and statistics.
Required Textbook and Other Course Material

• Joel Grus, Data Science from Scratch: First Principles with Python 1st Edition
• Matthew A. Russell, Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More
• Other material will be either provided by the lecturer or will be available online

Optional Reading

• Sandy Ryza, Uri Laserson, Sean Owen, and Josh Wills, Advanced Analytics with Spark
• Ian Goodfellow, Yoshua Bengio, and Aaron Courville, Deep Learning
• Doing Data Science

Prerequisites

• Comfortable working with or keenness to learn working with large set of data
• Comfortable with learning or relearning Calculus, Probability, Linear Algebra, and Random Variables
• Familiarity with and/or passion to learn a programming language and other computing tools. To be successful in computing (like many other disciplines) you need to persevere while solving your problems and learning.
• Bring a laptop to class in each lesson with necessary software installed.

Computation of Course Grade

The course grade will be based on the following:

• In-class tests: 35%
• Take-home assignments: 25%
• Team project: 30%
• Class participation: 10%

Course Policies

• You need to bring a valid form of picture ID on scheduled days of exams. You may not be allowed to take the exam without one.
• Communicating with others in class via electronic devices such as cellphones, tablets, and devices is strictly banned during class and during class tests. You will be reported for cheating if caught doing so.
• All grades are final when filed by the instructor on the Final Grade Report.
• You are required to bring a functioning laptop computer to the class that can effectively run Python environment (Anaconda Jupyter) and Tableau. Python is available under open source license. Academic license for Tableau will be provided. You will be allowed to use your laptop for some of tests but no communication with anyone will be accepted (see above).

Assignment Policies

• There will be multiple in-class tests to be held. There will be no make-up for in-class tests.
• In addition to class tests, there will home assignments, project assignments, class activities, and discussion questions.
• Home assignments made available in a class will be due by beginning of the following lesson. There will be no make-up for take-home assignments.
• There may be one or two extra-credit assignments or tests totaling no more than 5%. However, do not bank on extra-credit, as none may be available.
• The lecturer will randomly select project teams with each team consisting of 4 to 7 students. The lecturer in discussion with students will determine the topic for the project. The team project will be graded based on a mid-term presentation, a final presentation, a final submission (presentations, data sets, programs, and other relevant material), and by peer and cross-team rating.
• Class participation is dependent upon your attendance and your participation in the discussions in the class and in asking questions.
• Late and/or emailed solutions for projects or homework assignments will not be accepted. No exceptions will be made on this.
• All programming assignment must be submitted with the respective code that could run without any modification. If the code does not run, you may not get any point.
• Incomplete project and/or homework solutions (e.g. missing code), will not receive any credit.
• Homework and projects must be typed, unless there are analytical questions, in which case, the answers to these questions may be hand-written but should be neat and legible.
• You are advised to start working on the assignments as early as possible in order to have enough time to troubleshoot any issue(s) you may encounter. You are encouraged to ask question but asking basic questions at the 11th hour may not help you complete your work on time.
• You may work together on the take-home assignments with your class fellows. However, you must submit your own solution for each assignment. Plagiarism will result in zero credit for all involved.

**Suggested Paper Topics**

• Exchange Rate Prediction: Use Machine learning to predict currency exchange rate from multitude of data sources (historical data, social media, news, economic indicators, etc.).

• Sentiment Analysis for Economic Well-being of Tweeting Population: Devise sentiment analysis methods for assessing economic well-being of tweets.

• On prospects of using Deep Learning for Predicting Economic Meltdowns: Determine how Deep Learning and Artificial Neural Networks may be used to predict economic meltdowns.
Course Topics

The course is divided into 5 modules each spreading over two weeks and covered in two class sessions. Some topics may spill over their module especially where long lead-time is needed. Outline of the topics covered in modules is provided below. Next pages give more expanded detail of first Module 1 and Module 2. The other modules follow similar pattern:

Module 1: Introduction to Data Science and Exploratory Data Analysis
Introductions; Why Data Science; Data Science pipeline: Visualization, Exploratory Data Analysis (EDA), Data Modeling, and Production; Data Analytics and Big Data; Data Science and Data Scientist; Economics use cases in Data Science; Traits of a Data Scientist; Why Data Science is a multidisciplinary domain; Domain dependence of Data Science; Data Scientist is small and large organization; How Data Science is different than Statistics; EDA in Data Science and Statistics; Applications of Data Science.

Module 2: Machine Learning and Deep Learning
Introduction to Machine Learning; Supervised and unsupervised learning; Training and test data sets; Python programming language; Similarity measures and data normalization; Distance formula and its importance; Distance formula for fuzzy variables; Linear Regression and Naïve-Bayes with analytical basis; Python implementation of Linear Regression and Naïve-Bayes; Classification (N-Nearest Neighbor) and clustering (K-Means); Neural Networks and other Machine Learning algorithms; Machine Learning Services: Google TensorFlow for Machine Intelligence; Amazon AWS Machine Learning for Predictive Analytics; Skills for using cloud-based Machine Learning platforms.

Module 3: Big Data Processing
Data storage; Files systems and FATS; RDBMS and SQL; Limitations of traditional data analytics architecture; Schema-on-write and Schema-on-read; Hadoop - scalable, fault-tolerant, self-healing, distributed; Explosion of Data and IoT; 3 V's and 7 V's of Big Data; Data Centers; Hadoop as open source project; Hadoop Common; HDFS and MapReduce; Hadoop Ecosystem; Hadoop and MapReduce operations; Setting up a single-node Hadoop; Common Hadoop shell commands; YARN, Pig, Hive, and other tools; Hadoop distributions; NoSQL Databases; Using Amazon AWS DynamoDB for storing Tweets.

Module 4: Natural Language Processing and Sentiment Analysis
Natural Language Processing (NLP); Uses of NLP in Data Science; NLP and Machine Learning; Google Translate; Social media and Sentiment Analysis; Using NLP in product reviews and review filtering; Trending topics with NLP; Text block summarization; Topic extraction; Keyword topic tags generation using LDA (Latent Dirichlet Allocation); Auto-Tag and Auto-Tag URL micro-services; Sentiment Analysis tools to assess feelings and opinion; Sentiment analysis of Twitter feeds; Open Source NLP libraries.

Module 5: Data Visualization and Social Psychology in Data Science
Data Visualization (DV); Dimensionality of data in Data Science and abstraction; Temporal and other data; Terminology; Multivariate DV; Perception, colors and design; Text DV; Sparkle lines and other micro-level DV; Interactivity and animation; Geospatial DV; Tableau; Using Tableau for DV. Social Psychology and its uses in Data Science exploration: Human biases, Stereotyping of social applications, correlation vs. causality.
Details of First Two Modules
Details of the first two modules are provided below. Note that these details are provided only as guidelines as the actual content is subject to change.

Module 1: Introduction to Data Science and Exploratory Data Analysis

Mathematics Preliminaries:
**Linear Algebra**: Scalar, vectors, matrices, and tensors; determinants; multiplication and inverses; Eigen decomposition; SVD
**Probability and Information Theory**: Random variables, probability distributions, marginal and conditional probability, chain rule, Bayes’ rule, information theory.
**Numerical Computation**: Overflow and underflow, gradient-based optimization, constrained optimization, linear least squares.
**Calculus**: Functions, n-dimensional functions; derivatives and differentials; partial differentiation; system of linear equations.

Computing Preliminaries:
**Basics**: User of programming in data analytics; programming languages; Python programming; program development and IDEs
**Python variables and operations**: Strongly typed languages; numeric and strings; lists, tuples, dictionaries; multi-dimensional lists, creating lists; creating tokens
**Processing constructs**: Single and Doubly Nested Loops, Conditions, Booleans, Functions

Main Topics:
**Basics**: Introductions; what is Data Science and how it relates to Data Analytics, and Big Data. Data Science pipeline – from exploration to production; The EDA process; type of problem we address in Data Science; how economists use Data Science. Role of visualization, Exploratory Data Analysis (EDA), Data Modeling, and Production; Data Analytics and Big Data; Data Science and Data Scientist; what does a Data Scientist do; why Data Scientist need to know many things; traits of a Data Scientist; why Data Science is a multidisciplinary domain; domain dependence of Data Science; Data Scientist is small and large organization; How Data Science is different than Statistics; What is EDA in Data Science and Statistics;

Applications Examples: Recommendation systems; Sentiment analysis; Latent Semantic Analysis; etc.

Class Project: Social Media class project will be based on one of the popular social media API including Twitter, Facebook, LinkedIn, or Google +; project team will define the problem, gather data, program, and deliver presentations.

Assignments and Tests for Module 1:
1.1. Test: Preliminaries Quiz – 0 point
1.2. Assignment: Write Your Personal Data Science Statement – 4 point
1.3. Test: Python 1.A: Set up Python environment including Anaconda and PyCharm – 1 point
1.4. Test: Python 1.B: Variables, lists, tuples, dictionaries, operations, and range function – 1 point
1.5. Test: Python 1.C: Single and double nested loops, conditions, Booleans, functions – 2 point
1.6. Project: Define Social Media class project – Preliminary – 2 point
1.7. Project: Define Social Media class project – Final – 3 point
1.8. Test: Mathematical Preliminaries – 4 point
1.9. Class Activity: Define similarity measures for human resemblance – 1 point
1.10. Assignment: Calculate distance on a map – 2 points
Module 2: Machine Learning and Deep Learning

Python Topics: Write useful functions; returning multiple values; passing a variable list of parameters (*args and **kwargs); import and use libraries; numpy, pandas.

Machine Learning Basics Topics: Learning algorithms and their definition; the task, the performance measure, and the experience; capacity, over-fitting and under-fitting; supervised and unsupervised learning; training and test data sets; assumptions of similarity of data (i.i.d); similarity measures and data normalization; distance formula and it’s importance; distance formula for fuzzy variables

Main Topics:
Traditional Machine Learning: Linear Regression and Naïve-Bayes with analytical basis; Linear regression in open and matrix form; Logistic Regression; Decision Trees; Classification (N-Nearest Neighbor) and clustering (K-Means);

Deep Learning: The need for Deep Learning; a historical perspective on Deep Learning and Neural Networks; Multi-layered Perceptron (MLP) and Back-propagation; Other Deep Learning networks; Machine Learning Services: Google TensorFlow; Amazon AWS Machine Learning for Predictive Analytics; Skills for using cloud-based Machine Learning platforms.

Assignments and Tests for Module 2:
2.1. Assignment: Solve Linear Regression Analytically and Confirm Results – 3 point
2.2. Test: Python 2.A: Numpy and Pandas – 2 point
2.3. Test: Python 2.B: Tensorflow – 2 point
2.4. Assignment: Implement Linear Regression or Naïve Bayes Classifier in Python – 3 point
2.5. Project: Present Plan (data, goals, path) – 5 point
2.6. Class Activity: Define distance measure – 2 point
2.7. Discussion Question: Write an Economics Use Case for Machine Learning – 3 points